Common Errors in Marketing Experiments and How to Avoid Them

Tanya Kolosova, Samuel Berestizhevsky

Abstract

A methodology for designing experiments developed by Sir Ronald Fisher is more than 80 years old, but many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. Because marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not only inefficient but actually are not suitable for conducting marketing experiments.

In this article, we describe very useful and efficient split-unit (or split-plot) design of marketing experiments. Split-unit design is often used in marketing experiments but not recognized; often miss or inappropriately analyzed. We use a real-life example to demonstrate some of the ideas involved and ways to correctly analyze split-unit design.

Keywords: design of experiments, marketing experiments, split-unit design, split-plot design, bias correction, block variables, SAS.

Introduction

A methodology of design of experiments (DoE) was developed by Sir Ronald Fisher in his groundbreaking book “The Design of Experiments” yet in 1935. For his contribution in statistics, Sir Ronald Fisher has been described as "a genius who almost single-handedly created the foundations for modern statistical science" (Hald, 1998) and "the single most important figure in 20th-century statistics" (Efron, 1998). Since then, this methodology has been broadly adopted in the agricultural engineering, physical and social sciences, advertising and marketing.

Surprisingly, many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. There are multiple reasons to replace A/B tests by design of experiments:

a) In a design of experiments, the approach is completely different from A/B testing, as all parameters (factors) are changed together, simultaneously, and not one parameter at a time. Thus, in DoE the required number of experiments limited and significantly smaller than with A/B testing.
b) DoE provides a way to account for different sources of errors and compares averages to other averages rather than individual values to other individual values (as A/B testing). This allows achieving much greater accuracy in the estimation of factors effect for a given number of experiments, and thus the influential factors and their combinations are much more likely to emerge from the noise of the experimental errors.
c) But what is more critical, DoE allows estimating the impact of factors interactions which is not available in A/B testing. Because, in fact, marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not really suitable for conducting of marketing experiments.

DoE methodology creates a framework for planning, analyzing and executing marketing experiments. As with any statistical method, to receive correct results the method should be correctly applied.

One of the very efficient and often used designs is split-unit (also referred as split-plot) design, when one experimental unit is split into subunits, to which subsequent treatments are applied. Marketing usually involves a number of sequential steps, which makes split-unit design not only feasible and desirable but actually necessary.

The challenge is that split-unit experiments are often used but can be difficult to recognize. As a result, split-unit experiments often inappropriately analyzed. A spreadsheet of data can look like a variety of multifactor experiments, and it is very tempting to consider the experiment as completely randomized design (CRD) and then to apply straightforward analysis. In split-unit designs of experiments, it can take some research work to find out what factors, if any, are blocking factors and which are treatment factors, and, most importantly, what were the experimental units (EU) to which treatment factors were applied.

In the case of complex split-unit design, miss-interpretation of EU and incorrect error structure lead to inappropriate analysis, this, in turn, produces misleading results that may be very costly in marketing.

Description of the Marketing Experiment

As a real-life example, let's consider the case in which office supply retailing company A needs to test the impact of marketing emails to find out optimal combination of factors and achieve maximum sales as a response to marketing emails.

There are multiple factors which affect the success of email marketing, for example, factors that describe the marketing message, format the message was delivered, and audience that received the message.

In our real-life marketing experiment, the following 4 factors were included:

Factor Name	Levels	Factor Description
customer	C1 C2 C3	The company A differentiates their customers into 3 types according to customers purchasing behavior.
minimal_order	\$50 \$100	To become eligible for the discount, a customer has to make an order for a specific dollar value (at least).
discount	5% 10% 15%	If eligible according to the order dollar value, the customer will receive a discount on the whole order.
subject_line	SL1 SL2	2 versions of email subject lines were developed by the marketers for the marketing experiment

To quantify the success of the marketing experiment, the company A used total sales generated by the specific email marketing campaign.

First, lists of customers of 3 different types were created. The lists were created as a random selection from the repository of the company customers without returning the selected customer back to the repository. Customers were selected by customer types, producing 600,000 email recipients in each list. 4 replications of each type of customer were obtained, 12 Lists in total.

Then, each List was randomly divided into 6 Batches of 100,000 recipients, and the Batches were randomly assigned combinations of the minimal order value and discount: (\$50, 5%), (\$50, 10%), (\$50, 15%), (\$100, 5%), (\$100, 10%), (\$100, 15%).

Next, each Batch was randomly divided into 2 Groups of 50,000 recipients each. Each Group was randomly assigned SL1 or SL2 version of email subject line.

The table below (experimental table) presents the created full factorial experiment $2^2 3^2$ (36 treatment combinations) where each experiment cell contained 50,000 email recipients. The 4 replications of this experiment were conducted with an interval of 3 days.

Exp. run	customer	minimal_ order	discount	subject_ line
1.	C1	$50	5%	SL1
2.	C1	$50	5%	SL2
3.	C1	$50	10%	SL1
4.	C1	$50	10%	SL2
5.	C1	$50	15%	SL1
6.	C1	$50	15%	SL2
7.	C1	$100	5%	SL1
8.	C1	$100	5%	SL2
9.	C1	$100	10%	SL1
10.	C1	$100	10%	SL2
11.	C1	$100	15%	SL1
12.	C1	$100	15%	SL2
13.	C2	$50	5%	SL1
14.	C2	$50	5%	SL2
15.	C2	$50	10%	SL1
16.	C2	$50	10%	SL2
17.	C2	$50	15%	SL1
18.	C2	$50	15%	SL2
19.	C2	$100	5%	SL1
20.	C2	$100	5%	SL2
21.	C2	$100	10%	SL1
22.	C2	$100	10%	SL2
23.	C2	$100	15%	SL1
24.	C2	$100	15%	SL2
25.	C3	$50	5%	SL1
26.	C3	$50	5%	SL2
27.	C3	$50	10%	SL1
28.	C3	$50	10%	SL2
29.	C3	$50	15%	SL1
30.	C3	$50	15%	SL2
31.	C3	$100	5%	SL1
32.	C3	$100	5%	SL2
33.	C3	$100	10%	SL1
34.	C3	$100	10%	SL2
35.	C3	$100	15%	SL1
36.	C3	$100	15%	SL2

How Analysis Was Performed

This experiment was considered by company A as completely randomized design (CRD) and analyzed as such. The analysis was performed using SAS® Software.

```
proc mixed data=experiment cl;
class replication customer minimal_order discount subject_line;
model sales=customer|minimal_order|discount|subject_line;
run;
```

The randomization structure of the CRD implies that there is only one error term (the within error) and all factors effects are tested against it.

The results are presented in the table below:

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	108	208.37	<.0001
minimal_order	1	108	0.57	0.4525
customer*minimal_order	2	108	2.08	0.1304
discount	2	108	10.65	<.0001
customer*discount	4	108	5.22	0.0007
minimal_order*discount	2	108	0.00	0.9956
customer*minimal_order*discount	4	108	1.29	0.2784
subject_line	1	108	1.61	0.2072
customer*subject_line	2	108	2.70	0.0717
minimal_order*subject_line	1	108	9.89	0.0021
customer*minimal_order*subject_line	2	108	0.69	0.5059
discount*subject_line	2	108	3.42	0.0364
customer*discount*subject_line	4	108	3.11	0.0183
minimal_order*discount*subject_line	2	108	2.53	0.0843
customer*minimal_order*discount*subject_line	4	108	2.17	0.0767

Significant (on 95% confidence level) factors and their interactions are customer, discount, customer*discount, minimal_order*subject_line, discount*subject_line and customer*discount*subject_line. Using these factors, we built regression model and found conditions (factors and their levels) that maximize response (sales).

```
proc mixed data=experiment cl;
class replication customer min_order discount subject_line;
model sales=customer discount customer*discount customer*subject_line
            minimal_order*subject_line customer*discount*subject_line
      /solution singular=1e-11 ddfm=kr outpm=pred;
run;
```

For each customer type, the conditions generating maximum sales presented in the table below:

customer	minimal_order	discount	subject_line	Predicted sales
C1	$50	15%	SL1	$130,681
C2	$50	10%	SL1	$168,058
C3	$100	15%	SL2	$179,607

How Analysis Should Be Performed

We suggest a closer look at how the experiment was executed to understand if the analysis of the experiment was performed correctly.

First, customers were randomly selected by customer types, producing 600,000 email recipients in 12 Lists: 4 replications of each of 3 types of customers. This created a completely randomized design. The list was an experimental unit (EU) for types of customers (3 levels) – the entity to which types of customers are randomly assigned (see Figure 1).

Then, each List was randomly divided to 6 Batches of 100,000 recipients, with randomly assigned combinations of the minimal order condition and percent of discount: ($50, 5%), ($50, 10%), ($50, 15%), ($100, 5%), ($100, 10%), ($100, 15%). The act of grouping the experimental units together into homogenous groups is called blocking. Thus, List was a block of 6 Batches, and Batch was an experimental unit for combinations of the minimum order and discount. In other words, Batch design is a randomized complete block design, where List is the blocking factor (see Figure 2).

And when each Batch was randomly divided into 2 Groups for 2 versions of email subject lines, Batch (List) was a block for levels of email subject lines (see Figure 3).

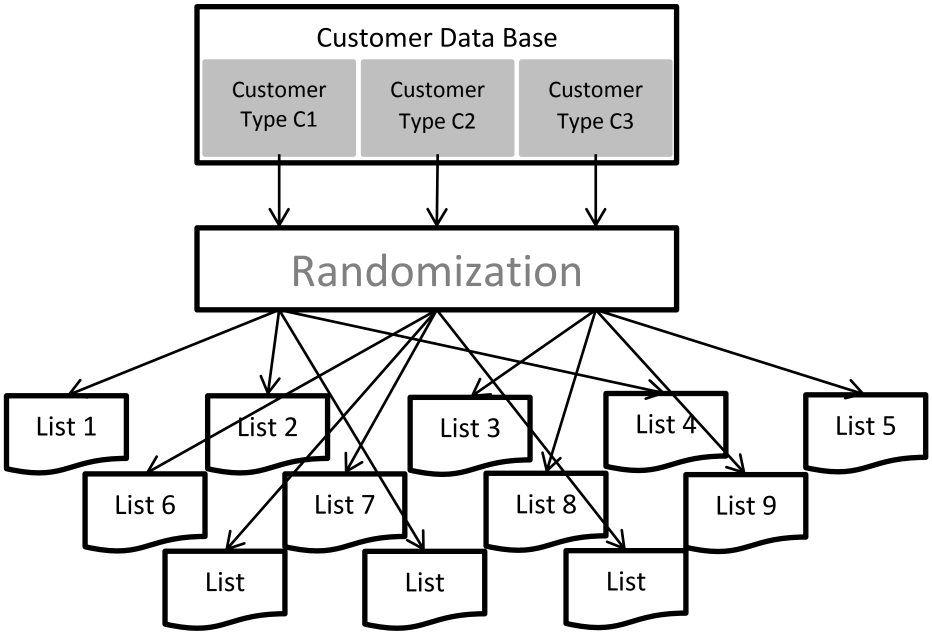

Figure 1. Lists Randomization

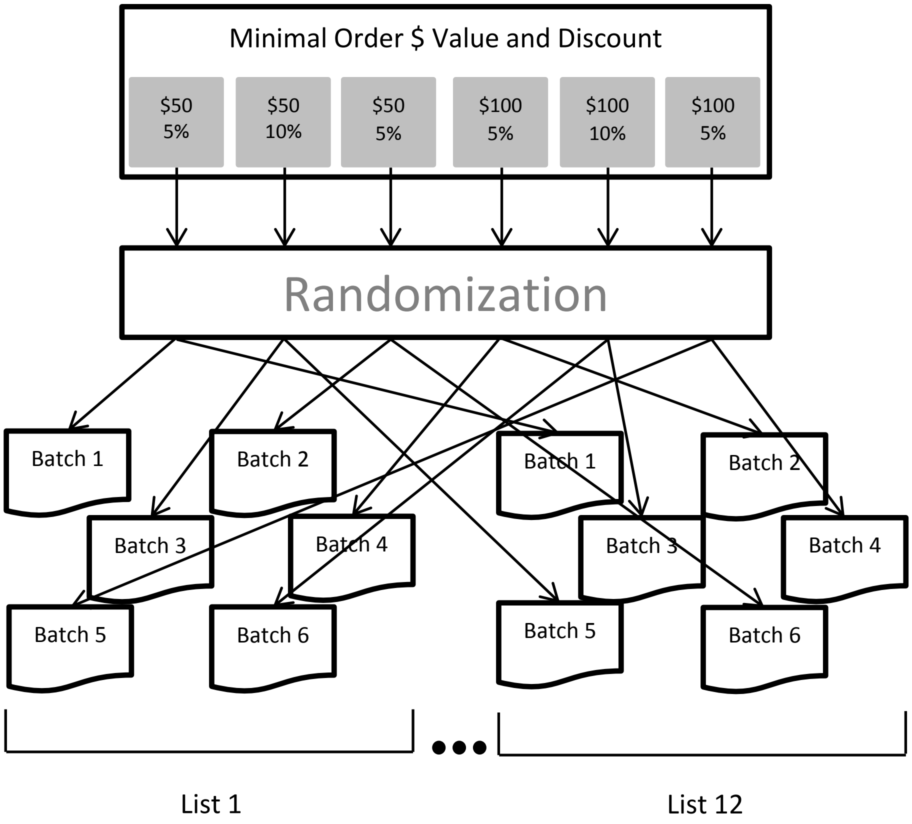

Figure 2. Batch Randomization

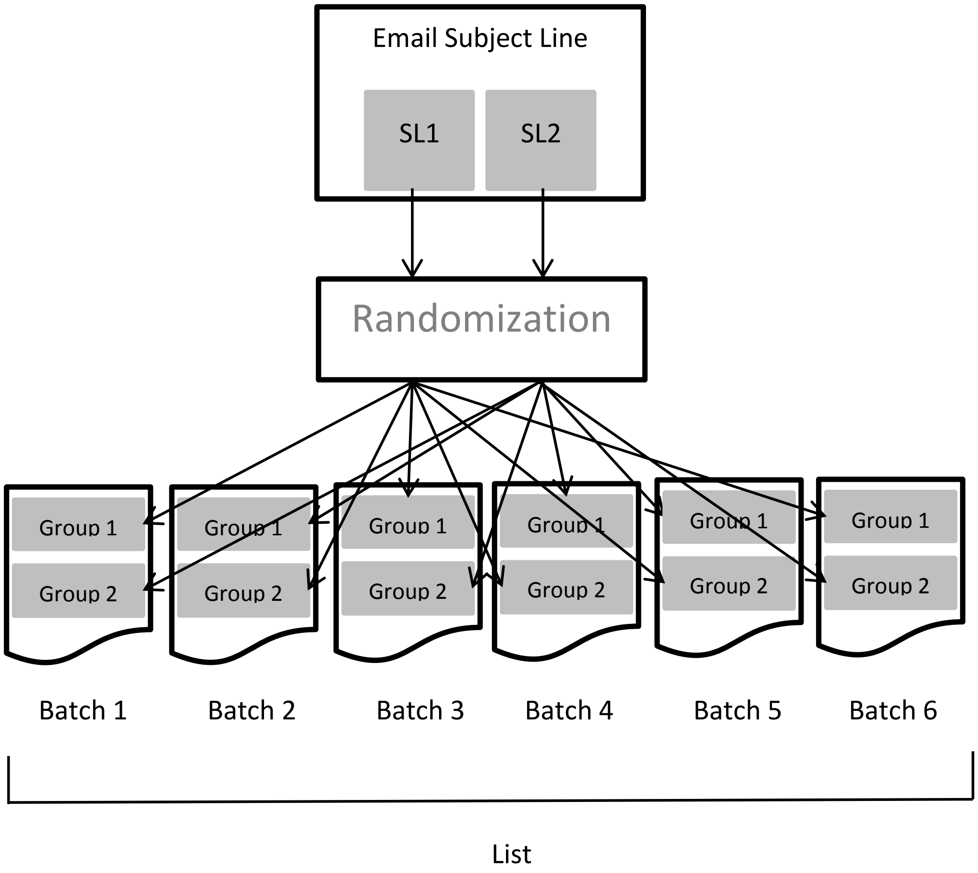

Figure 3. Groups Randomization

As a result, the appropriate model should include:

- Factorial effects for levels of customer type * minimal order * offered discount * email subject line,
- and 3 sizes of experimental units: List, Batch, Group

Using split-unit error structure, we analyzed the results of the same experiment.

```
proc mixed data=experiment cl;
class replication customer minimal_order discount subject_line;
model sales=customer|minimal_order|discount|subject_line;
random replication(customer) minimal_order*discount*replication(customer);
run;
```

Results of the analysis are presented below:

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	9	47.75	<.0001
minimal_order	1	45	0.62	0.4362
customer*minimal_order	2	45	2.25	0.1166
discount	2	45	11.56	<.0001
customer*discount	4	45	5.67	0.0009
minimal_order*discount	2	45	0.00	0.9953
customer*minimal_order*discount	4	45	1.40	0.2490
subject_line	1	54	34.86	<.0001
customer*subject_line	2	54	40.97	<.0001

minimal_order*subject_line	1	54	1.57	0.2155
customer*minimal_order*subject_line	2	54	1.36	0.2660
discount*subject_line	2	54	0.77	0.4681
customer*discount*subject_line	4	54	6.15	0.0004
minimal_order*discount*subject_line	2	54	5.01	0.0101
customer*minimal_order*discount*subject_line	4	54	4.30	0.0043

Significant factors and interactions were customer, discount, subject_line, customer*discount, customer*subject_line, customer*discount*subject_line, minimal_order*discount*subject_line and customer*minimal_order*discount*subject_line.

Now, we built a new regression model, and estimated conditions that maximized response (sales).

```
proc mixed data=experiment cl;
class replication customer minimal_order discount subject_line;
model sales=customer discount subject_line customer*discount customer*subject_line
            customer*discount*subject_line minimal_order*discount*subject_line
            customer*minimal_order*discount*subject_line
      /solution singular=1e-11 ddfm=kr outpm=pred_split;
random replication(customer) minimal_order*discount*rep(customer);
run;
```

For each customer type, the conditions generating maximum sales presented in the table below:

customer	minimal_order	discount	subject_line	Predicted sales
C1	$50	10%	SL1	$140,174
C2	$100	10%	SL1	$156,830
C3	$100	10%	SL2	$191,097

Comparison of the Results

Split-unit error structure allowed to discover different interactions that existed in the experimental data. The reason is that CRD analysis pools the three error terms together and the resulting error is not appropriate for any of the comparisons. In fact, the split-unit design is more complex, and it has more relationships among factors than discovered using CRD.

CRD analysis found the interactions minimal_order*subject_line and discount*subject_line significant, while split-unit didn't. On the other hand, split-unit found subject_line factor and interactions customer*subject_line, minimal_order*discount*subject_line and customer*minimal_order*discount*subject_line significant, while CRD did not.

As a result, CRD analysis identified incorrectly the conditions generating a maximum response(sales), and what response (sales) can be actually achieved.

What impact would it have on the actual business performance?

According to the CRD analysis, the best conditions for customer type C1 are 15% discount with minimum purchase of $50 while email is sent with subject line SL1. These conditions should bring $130,681 in sales per 50,000 recipients. However, if we substitute these conditions into the model built based on the split-unit analysis, the result will be $127,094 – 2.7% less. If the campaign would be sent to 1,000,000 recipients it would translate to about $71,000 lower sales than expected.

For the same type of customers, the split-unit analysis identified conditions of 10% discount with minimum purchase of $50 while email is sent with subject line SL1. Under these conditions, the expected sales from 50,000 of email recipients are $140,174. In comparison with $127,094 that would be received under conditions identified by CRD analysis, the correct conditions would generate 10.3%

more sales. And if the marketing emails with the conditions identified by split-unit analysis would be sent to 1,000,000 recipients it would translate to $261,600 higher sales.

When we perform a similar examination for customer type C2, the results are the following:

- CRD analysis suggests that the best conditions (10%, $50, SL1) will generate $168,058.
- If we plug in these conditions into the split-unit model, the expected sales are $155,070, which is 7.73% less. Applied to a campaign for 1,000,000 recipients this will produce $259,760 less than expected.
- The split-unit analysis suggests that the best conditions (10%, $100, SL1) will generate $156,830. For 1,000,000 recipients this will produce $35,200 more than based on the conditions identified by CRD analysis.

For customer type C3, the results are the following:

- CRD analysis suggests that the best conditions (15%, $100, SL2) will generate $179,607.
- If we plug in these conditions into the split-unit model, the expected sales are $168,914, 5.95% less. Applied to a campaign for 1,000,000 recipients this will produce $213,860 less than expected.
- The split-unit analysis suggests that the best conditions (10%, $100, SL2) will generate $191,097. For 1,000,000 recipients this will produce $443,660 more than based on the CRD conditions.

Summary

Design of experiment applied to marketing helps identifying factors and their interactions that maximize marketing campaigns performance.

Failure to identify the appropriate design structure leads to an incorrect analysis of the experiment, and as a result, produces misleading inferences.

We demonstrated that incorporating the split-unit error structure provides appropriate analyses, comparisons, and correct predictive models.

References

Box, G.E.P., Hunter, W.G., Hunter, J.S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York: Wiley.

Efron, B. (1998). R. A. Fisher in the 21st century. Statistical Science, 13: 95–122.

Hald, A. (1988). A History of Mathematical Statistics. New York: Wiley.

Kolosova, T., Berestizhevsky, S. (1998). Programming Techniques for Object-Based Statistical Analysis with SAS Software. Cary, NC: SAS Institute Inc.

Littell, R.C.L., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

Wu, C.F.J, Hamada, M. (2000). Experiments: Planning, Analysis, and Parameter Design Optimization. New York: Wiley.